REPORT DOCUMENTATION PAGE

AFRL-SR-AR-TR-08-0034

1. REPORT DATE *(DD-MM-YYYY)*
12/3/2007

2. REPORT TYPE
Final Technical Report

3. DATES COVERED *(From - To)*
3/1/04-8/31/07

4. TITLE AND SUBTITLE
Algorithms for Data Sharing, Coordination, and Communication in Dynamic Network Settings

5a. CONTRACT NUMBER
FA9550-04-1-0121

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Prof. Nancy Lynch

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Dr. Robert L. Herklotz
Program Manager, Software and Systems
Air Force Office of Scientific Research
875 Randolph St., Suite 325, Room 3112
Arlington, VA 22203-1768

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

No limits

Approved for public release. Distribution is unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This project developed many distributed algorithms and corresponding lower bounds for solving important problems in dynamic networks, focusing on mobile networks with wireless communication. Problems studied include data management, time synchronization, communication problems (broadcast, geocast, and point-to-point routing), distributed consensus, tracking, and motion coordination. Highlights include (1) The discovery of a fundamental limitation in capabilities for time synchronization in large networks. (2) The identification and development of the notion of ``Virtual Node Layers" as abstraction layers for programming mobile networks; these appear to facilitate programming of key communication services, as well as motion coordination for robots, vehicles and aircraft. (3) Upper and lower bounds for solving basic problems such as distributed consensus in mobile networks in which messages are subject to loss and collisions.(4) The development of a mathematical framework---a combination of Timed and Probabilistic I/O Automata---capable of modeling the dynamic networks and algorithms that were studied, and of supporting theorems about correctness and performance of the algorithms.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:
unclassified

17. LIMITATION OF ABSTRACT

18. NUMBER OF PAGES

19a. NAME OF RESPONSIBLE PERSON
Nancy Lynch

a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(include area code)* 617-253-7225

Final Technical Report for USAF,AFRL Award #FA9550-04-1-0121,
Algorithms for Data Sharing, Coordination, and Communication in Dynamic Network Settings
March 1, 2004 - August 31, 2007
PI Nancy Lynch

1 Abstract

This project developed many distributed algorithms and corresponding lower bounds for solving important problems in dynamic networks, focusing on mobile networks with wireless communication. Problems studied include data management, time synchronization, communication problems (broadcast, geocast, and point-to-point routing), distributed consensus, tracking, and motion coordination. Highlights include (1) The discovery of a fundamental limitation in capabilities for time synchronization in large networks. (2) The identification and development of the notion of "Virtual Node Layers" as abstraction layers for programming mobile networks; these appear to facilitate programming of key communication services, as well as motion coordination for robots, vehicles and aircraft. (3) Upper and lower bounds for solving basic problems such as distributed consensus in mobile networks in which messages are subject to loss and collisions. (4) The development of a mathematical framework—a combination of Timed and Probabilistic I/O Automata—capable of modeling the dynamic networks and algorithms that were studied, and of supporting theorems about correctness and performance of the algorithms.

2 People

Faculty:
Nancy Lynch

Postdoctoral Associates:
Ling Cheung
Gregory Chockler
Murat Demirbas
Dilsun Kirli Kaynar

PhD Students:
Rui Fan
Seth Gilbert
Sayan Mitra
Calvin Newport
Tina Nolte
Joshua Tauber
Shinya Umeno

MEng students and Undergraduate Researchers:
Matthew Brown
Catherine Matlon
Mike Spindel

Principle Collaborators:
Paul Attie (Northeastern University)
Shlomi Dolev (Ben-Gurion University)
Ralph Droms (Cisco)
Nancy Griffeth (Lehman College, CUNY)
Rachid Guerraoui (EPF Lausanne)
Roberto Segala (University of Verona)
Alex Shvartsman (University of Connecticut)
Frits Vaandrager (University of Nijmegen)
Jennifer Welch (Texas A&M University)

3 Theses

Matt Brown (Meng thesis)
Seth Gilbert (PhD thesis)
Catherine Matlon (Meng thesis)
Sayan Mitra (PhD thesis)
Calvin Newport (MS thesis)
Shinya Umeno (MS thesis)

4 Progress, with Publications and Abstracts

Our overall project was designed to develop distributed algorithms and corresponding lower bounds for solving important problems in dynamic networks. "Dynamic networks" here include mobile networks, sensor nets, client-server networks, peer-to-peer networks, etc. In such settings, participants may join, leave, fail, or move (either controllably or uncontrollably).

In these settings, we have studied a wide range of problems, including coherent data management, time synchronization, communication problems (broadcast, geocast, point-to-point routing), Virtual Node emulation, distributed consensus, location management, tracking, resource allocation, and motion coordination.

Along the way, we have developed necessary mathematical foundations, including models, complexity measures, and analysis methods. In general, we model systems in terms of interacting automata—mainly, I/O Automata, Timed I/O Automata, Hybrid I/O Automata, and Probabilistic I/O Automata.

Motivating applications for this work are communication systems (both traditional and mobile ad hoc), controlled systems (of robots, vehicles, and aircraft), and to a certain extent, security protocols. Our recent work on security protocol modeling and analysis has been supported primarily by an NSF ITR grant with Micali and Rivest.

Our mathematical foundations work has been complemented by work on the TIOA (Tempo) Toolset. This work, however, has been conducted under the support of other grants from AFOSR and NSF, so I am not describing it in this final report.

The subsections below include references to the various papers we have produced on this contract, together with their abstracts and a running overview and commentary.

4.1 Basic Algorithms for Dynamic Networks

Our initial work on dynamic network algorithms involved study of basic problems of reliable broadcast, time synchronization, and location management (tracking).

4.1.1 Reliable broadcast

Carolos Livadas and Nancy A. Lynch. A Reliable Broadcast Scheme for Sensor Networks. Technical Report MIT-LCS-TR-915, MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, February 2007. (Revision of earlier version dated August 2003).

Abstract: In this short technical report, we present a simple yet effective reliable broadcast protocol for sensor networks. This protocol disseminates packets throughout the sensor network by flooding and recovers from losses resulting from collisions by having hosts retransmit packets whenever they notice that their neighbors have fallen behind. Such retransmissions serve to flood the appropriate packets throughout the regions of the sensor network that did not receive the given packets as a result of prior flooding attempts.

4.1.2 Time synchronization

Standard theoretical time synchronization algorithms were designed for fixed, known networks; these algorithms do not extend to dynamic, unknown networks. We carried out two projects on time synchronization, one involving design of a new time synch algorithms for dynamic networks, and the other observing an inherent limitation on the scalability of time synchronization. The second of these papers won a "Best Student Paper" award at PODC 2004. A focus of both papers is a new "gradient" property, which requires that neighboring nodes' clocks be closely synchronized.

Rui Fan, Indraneel Chakraborty, and Nancy Lynch. Clock Synchronization for Wireless Networks. In Teruo Higashino, editor, Principles of Distributed Systems: OPODIS 2004: 8th International Conference on Principles of Distributed Systems, Grenoble, France, December 15-17, 2004, volume 3544 of Lecture Notes in Computer Science, pages 400-414, 2005. Springer.

Abstract: Time synchronization is a fundamental service in many wireless applications. While the synchronization problem is well-studied in traditional wired networks, physical constraints of the wireless medium impose a unique set of challenges. We present a novel time synchronization algorithm which is highly energy efficient and failure/recovery-tolerant. Our algorithm allows nodes to synchronize to sources of real time such as GPS when such signals are available, but continues to synchronize nodes to each other, even in the absence of GPS. In addition, the algorithm satisfies a relaxed gradient property, in which the degree of synchronization between nodes varies as a linear function of their distance. Thus, nearby nodes are highly synchronized, which is desirable in many wireless applications.

Rui Fan and Nancy Lynch. Gradient Clock Synchronization. Proceedings of the Twenty-Third Annual ACM SIGACT-SIGOPS Symposium on Principles of Distributed Computing (PODC 2004)), St. John's, Newfoundland, Canada, pages 320-327, July 25-58, 2004. Best Student Paper Award.

Rui Fan and Nancy Lynch. Gradient Clock Synchronization. Distributed Computing, volume 18, number 4, pages 255-266, March, 2006.

Abstract: We introduce the distributed gradient clock synchronization problem. As in traditional distributed clock synchronization, we consider a network of nodes equipped with hardware clocks with bounded drift. Nodes compute logical clock values based on their hardware clocks and message exchanges, and the goal is to synchronize the nodes' logical clocks as closely as possible, while satisfying certain validity conditions. The new feature of gradient clock synchronization (GCS for short) is to require that the skew between any two nodes' logical clocks be bounded by a nondecreasing function of the uncertainty in message delay (call this the distance) between the two nodes, and other network parameters. That is, we require nearby nodes to be closely synchronized, and allow faraway nodes to be more loosely synchronized. We contrast GCS with traditional clock synchronization, and discuss several practical motivations for GCS, mostly arising in sensor and ad-hoc networks. Our main result is that the worst case clock skew between two nodes at distance d or less from each other is π $(d + logDloglogD)$, where D is the diameter of the network. This means that clock synchronization is not a local property, in the sense that the clock skew between two nodes depends not only on the distance between the nodes, but also on the size of the network. Our lower bound implies, for example, that the TDMA protocol with a fixed slot granularity will fail as the network grows, even if the maximum degree of each node stays constant.

4.1.3 Location services (tracking)

We conducted some work on sensor network algorithms, partly under the auspices of the DARPA NEST project. The NEST project was focused on tracking of intruders, which led to our work on intruder tracking algorithms. We followed up this early work with a new type of intruder tracking algorithm using Virtual Nodes—see Section 4.3.3.

Murat Demirbas, Anish Arora, Tina Nolte, and Nancy Lynch. A Hierarchy-based Fault-local Stabilizing Algorithm for Tracking in Sensor Networks. In Teruo Higashino, editor, Principles of Distributed Systems: OPODIS 2004: 8th International Conference on Principles of Distributed

Systems, Grenoble, France, December 15-17, 2004, volume 3544 of Lecture Notes in Computer Science, pages 299-315, 2005. Springer.

Abstract: In this paper, we introduce the concept of hierarchy-based fault-local stabilization and a novel self-healing/fault-containment technique and apply them in STALK. STALK is an algorithm for tracking in sensor networks that maintains a data structure on top of an underlying hierarchical partitioning of the network. Starting from an arbitrarily corrupted state, STALK satisfies its specification within time and communication cost proportional to the size of the faulty region, defined in terms of levels of the hierarchy where faults have occurred. This local stabilization is achieved by slowing propagation of information as the levels of the hierarchy underlying STALK increase, enabling more recent information propagated by lower levels to override misinformation at higher levels before the misinformation is propagated more than a constant number of levels. In addition, this stabilization is achieved without reducing the efficiency or availability of the data structure when faults don't occur: 1) Operations to *find* the mobile object distance d away take $O(d)$ time and communication to complete, 2) Updates to the tracking structure after the object has moved a total of d distance take $O(dlog$ network diameter) amortized time and communication to complete, 3) The tracked object may relocate without waiting for STALK to complete updates resulting from prior moves, and 4) The mobile object can move while a *find* is in progress.

Murat Demirbas, Anish Arora, Tina Nolte, and Nancy Lynch. Brief Announcement: STALK: A Self-Stabilizing Hierarchical Tracking Service for Sensor Networks. Proceedings of the 23rd Annual ACM SIGACT-SIGOPS Symposium on Principles of Distributed Computing (PODC 2004), St. John's, Newfoundland, Canada, page 378, July 25-28, 2004.

We present STALK, a hierarchy-based fault-local stabilizing algorithm for tracking in sensor networks. Starting from an arbitrarily corrupted state, STALK satisfies its specification within time and communication cost proportional to the size of the faulty region instead of the network size. Local stabilization is achieved by slowing propagation of information as the levels of the hierarchy underlying STALK increase, enabling the more recent information propagated by lower levels to override misinformation at higher levels. While achieving fault-local stabilization, STALK also adheres to the locality of tracking operations: an operation to find a mobile object at a distance d away requires $O(d)$ amount of time and communication cost to intercept the moving object, and a move of an object to a distance d away requires $0(d * log(D))$ time and communication cost to update the tracking structure. Furthermore, STALK achieves seamless tracking of a continuously moving object by enabling concurrent executions of move and find operations.

4.1.4 Other basic algorithms

Sayan Mitra and Jesse Rabek. Energy Efficient Connected Clusters for Mobile Ad Hoc Networks. MED-HOC-NET 2004, Third Annual Mediterranean Ad Hoc Networking Workshop, Bodrum, Turkey, June 2004.

Abstract: A Mobile Ad Hoc Network (MANET) is a wireless infrastuctureless network with mobile nodes. Clustering is a common basis for building higher level applications for such networks. The merit of a clustered decomposition depends on the application that is meant to use it. A power control based distributed clustering service is proposed that maintains cluster connectivity under reasonable assumptions. The size and sparsity of the clustering can be controlled by two parameters, namely, the minimal separation between the clusterheads, and the maximum angular gap between neighboring clusterheads. The optimal value of the latter is derived; this minimizes the transmission power of the clusterheads while guaranteeing connectivity of the cluster graph. Experimental studies presented show that the algorithm rapidly stabilizes to a new clustered organization after the network topology changes due to node joins and failures.

Ben Leong, Sayan Mitra, and Barbara Liskov. Path vector face routing: Geographic routing with local face information. Proceedings of 13th IEEE International Conference on Network Protocols (ICNP'05), November 6-9, 2005, Boston, Massachusetts, USA.

Abstract: Existing geographic routing algorithms depend on the planarization of the network connectivity graph for correctness, and the planarization process gives rise to a well-defined notion of "faces". In this paper, we demonstrate that we can improve routing performance by storing a small amount of local face information at each node. We present a protocol, Path Vector Exchange Protocol (PVEX), that maintains local face information at each node efficiently, and a new geographic routing algorithm, Greedy Path Vector Face Routing (GPVFR), that achieves better routing performance in terms of both path stretch and hop stretch than existing geographic routing algorithms by exploiting available local face information. Our simulations demonstrate that GPVFR/PVEX achieves significantly reduced path and hop stretch than Greedy Perimeter Stateless Routing (GPSR) and somewhat better performance than Greedy Other Adaptive Face Routing (GOAFR+) over a wide range of network topologies. The cost of this improved performance is a small amount of additional storage, and the bandwidth required for our algorithm is comparable to GPSR and GOAFR+ in quasi-static networks.

4.2 Rambo algorithms

We developed our Rambo algorithms in part before the start of this AFOSR contract, and finished the work during the contract. Basically, the Rambo algorithms are designed to maintain simple read/write data objects in a changing network, for example, a mobile ad hoc network for soldiers or first responders. The first paper below describes an initial version by Lynch and Shvartsman; the second, which we call Rambo II, includes a very important performance optimization developed by Gilbert. Later projects involved engineering improvements, implementations, and applications.

4.2.1 Basic Rambo algorithms

Nancy Lynch and Alex Shvartsman. RAMBO: A Reconfigurable Atomic Memory Service for Dynamic Networks. In D. Malkhi, editor, Distributed Computing (Proceedings of the 16th International Symposium on DIStributed Computing (DISC) October 2002, Toulouse, France), volume 2508 of Lecture Notes in Computer Science, pages 173-190, 2002. Springer-Verlag. Also, Technical

Report MIT Laboratory for Computer Science, Technical Report MIT-LCS-TR-856, Cambridge, MA, 2002.

Seth Gilbert. RAMBO II: Rapidly Reconfigurable Atomic Memory for Dynamic Networks. Masters Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, August 2003. .

Seth Gilbert, Nancy Lynch, and Alex Shvartsman. RAMBO II: Rapidly Reconfigurable Atomic Memory for Dynamic Networks. Proceedings of the International Conference on Dependable Systems and Networks (DSN), San Francisco, CA, pages 259-268, June 22nd - 25th, 2003.

Seth Gilbert, Nancy Lynch, and Alex Shvartsman. RAMBO II: Implementing atomic memory in dynamic networks, using an aggressive reconfiguration strategy. Technical Report MIT-CSAIL-TR-890, CSAIL, Massachusetts Institute Technology, Cambridge, MA, 2004.

Abstract: This paper presents a new algorithm implementing reconfigurable atomic read/write memory for highly dynamic environments. The original Rambo algorithm, recently developed by Lynch and Shvartsman, guarantees atomicity for arbitrary patterns of asynchrony, message loss, and node crashes. Rambo II implements a different approach to establishing new configurations: instead of operating sequentially, the new algorithm reconfigures aggressively, transferring information from old configurations to new configurations in parallel. This improvement substantially reduces the time to establish a new configuration and to remove obsolete configurations. This, in turn, substantially increases fault tolerance and reduces the latency of read/write operations when the network is unstable or reconfiguration is bursty. This paper presents Rambo II, a correctness proof, and a conditional analysis of its performance. Preliminary empirical studies illustrate the advantages of the new algorithm.

4.2.2 Engineering improvements and implementations

Chryssis Georgiou, Peter M. Musial, and Alexander A. Shvartsman. Long-Lived Rambo: Trading Knowledge for Communication. Rastislav Kralovic, Ondrej Sykora (Eds.): Structural Information and Communication Complexity, 11th International Colloquium, SIROCCO 2004, Smolenice Castle, Slowakia, June 21-23, 2004, volume 3104 of Lecture Notes in Computer Science, pages 185-196, Springer 2004.

Chryssis Georgiou, Peter M. Musial, and Alexander A. Shvartsman. Long-Lived Rambo: Trading Knowledge for Communication. Technical Report MIT-LCS-TR-943, MIT CSAIL, Cambridge, MA, April 2004.

Abstract: Shareable data services providing consistency guarantees, such as atomicity (linearizability), make building distributed systems easier. However, combining linearizability with efficiency in practical algorithms is difficult. A reconfigurable linearizable data service, called Rambo, was developed by Lynch and Shvartsman. This service guarantees consistency under dynamic conditions involving asynchrony, message loss, node crashes, and new node arrivals. The specification of the original algorithm is given at an abstract level aimed at concise presentation and formal reasoning about correctness. The algorithm propagates information by means of gossip messages. If the service is in use for a long time, the size and the number of gossip messages may grow without bound. This paper presents a consistent data service for long-lived objects that improves on Rambo in two ways: it includes an incremental communication protocol and a leave service. The new protocol takes advantage of the local knowledge, and carefully manages the size of messages by removing redundant information, while the leave service allows the nodes to leave the system gracefully. The new algorithm is formally proved correct by forward simulation using levels of abstraction. An experimental implementation of the system was developed for networks-of-workstations. The paper also includes selected analytical and preliminary empirical results that illustrate the advantages of the new algorithm.

P.M. Musial and A.A. Shvartsman. Implementing a Reconfigurable Atomic Memory Service for Dynamic Networks. In Proceedings of the 18'th International Parallel and Distributed Processing Symposium (IPDPS'04) — FTPDS WS, Santa Fe, New Mexico, pages 208-215, April, 2004.

Abstract: Transforming abstract algorithm specifications into executable code is an error-prone process in the absence of sophisticated compilers that can automatically translate such specifications into the target distributed system. This paper presents a framework that was developed for translating algorithms specified as Input/Output Automata (IOA) to distributed programs. The framework consists of a methodology that guides the software development process and a core set of functions needed in target implementations that reduce unnecessary software development. As a proof of concept, this work also presents a distributed implementation of a reconfigurable atomic memory service for dynamic networks. The service emulates atomic read/write shared objects in the dynamic setting where processors can arbitrarily crash, or join and leave the computation. The algorithm implementing the service is given in terms of IOA. The system is implemented in Java and runs on a network of workstations. Empirical data illustrates the behavior of the system.

Gregory Chockler, Seth Gilbert, Vincent Gramoli, Peter Musial, and Alexander Shvartsman. Reconfigurable Distributed Storage for Dynamic Networks. 9th International Conference on Principles of Distributed Systems (OPODIS 2005), Pisa, Italy, December 12-14, 2005.

Abstract: This paper presents a new algorithm, RDS (Reconfigurable Distributed Storage), for implementing a reconfigurable distributed shared memory in an asynchronous dynamic network. The algorithm guarantees atomic consistency (linearizability) in all executions in the presence of arbitrary crash failures of processors and message loss and delays. The algorithm incorporates a quorum-based read/write algorithm and an optimized consensus protocol, based on Paxos. RDS achieves the design goals of: (i) allowing read and write operations to complete rapidly, and (ii) providing long-term fault tolerance through reconfiguration, a process that evolves the quorum configurations used by the read and write operations. The new algorithm improves on previously developed alternatives by using a more efficient reconfiguration protocol, thus guaranteeing better fault tolerance and faster recovery from network instability. This paper presents RDS, a formal proof of correctness, conditional performance analysis, and experimental results.

K. Konwar, P.M. Musial, N.C. Nicolau, and A.A. Shvartsman. Implementing Atomic Data through Indirect Learning in Dynamic Networks. Technical Report MIT-CSAIL-TR-2006-070, Computer

Science and Artificial Intelligence Laboratory, Massahusetts instute of Technology, Cambridge, MA, October 2006.

Abstract: Developing middleware services for dynamic distributed systems, e.g., ad-hoc networks, is a challenging task given that such services must deal with communicating devices that may join and leave the system, and fail or experience arbitrary delays. Algorithms developed for static settings are often not usable in dynamic settings because they rely on (logical) all-to-all connectivity or assume underlying routing protocols, which may be unfeasible in highly dynamic settings. This paper explores the indirect learning approach to information dissemination within a dynamic distributed data service. The indirect learning scheme is used to improve the liveness of the atomic read/write object service in the settings with uncertain connectivity. The service is formally proved to be correct, i.e., the atomicity of the objects is guaranteed in all executions. Conditional analysis of the performance of the new service is presented. This analysis has the potential of being generalized to other similar dynamic algorithms. Under the assumption that the network is connected, and assuming reasonable timing conditions, the bounds on the duration of the read/write operations of the new service are calculated. Finally, the paper proposes a deployment strategy where indirect learning leads to an improvement in communication costs relative to a previous solution.

4.2.3 Extensions and applications

Jacob Beal and Seth Gilbert. RamboNodes for the Metropolitan Ad Hoc Network. Proceedings of the Workshop on Dependability in Wireless Ad Hoc Networks and Sensor Networks, part of the International Conference on Dependable Systems and Networks, Florence, Italy, June-July, 2004. Also, AI Memo: AIM-2003-027.

Abstract: We present an algorithm to store data robustly in a large, geographically distributed network. It depends on localized regions of data storage that move in response to changing conditions. For example, data may migrate away from failures or toward regions of high demand. The PersistentNode algorithm of Beal provides this service robustly, but with limited safety guarantees. We use the Rambo framework to transform PersistentNode into RamboNode, an algorithm that guarantees atomic consistency in exchange for increased cost and decreased liveness. A half-life analysis of RamboNode shows that it is robust against continuous low-rate failures. Finally, we provide experimental simulations for the algorithm on 2000 nodes, demonstrating how it services requests and examining how it responds to failures.

Athicha Muthitacharoen, Seth Gilbert, and Robert Morris. Etna: A Fault-tolerant Algorithm for Atomic Mutable DHT Data. Technical Report MIT-CSAIL-TR-2005-044, MIT CSAIL, Cambridge, MA, June 2005.

Abstract: This paper presents Etna, an algorithm for atomic reads and writes of replicated data stored in a distributed hash table. Etna correctly handles dynamically changing sets of replica hosts, and is optimized for reads, writes, and reconfiguration, in that order. Etna maintains a series of replica configurations as nodes in the system change, using new sets of replicas from the pool supplied by the distributed hash table system. It uses the Paxos protocol to ensure consensus on the members of each new configuration. For simplicity and performance, Etna serializes all reads and writes through a primary during the lifetime of each configuration. As a result, Etna completes read and write operations in only a single round from the primary. Experiments in an environment with high network delays show that Etna's read latency is determined by round-trip delay in the underlying network, while write and reconfiguration latency is determined by the transmission time required to send data to each replica. Etna's write latency is about the same as that of a non-atomic replicating DHT, and Etna's read latency is about twice that of a non-atomic DHT due to Etna assembling a quorum for every read.

4.3 Virtual Nodes

Our work on Virtual Nodes began near the start of this AFOSR contract and evolved to occupy a significant portion of our research effort. Virtual Node Layers provide an abstraction layer for programming mobile ad hoc networks, such as networks of soldiers, rescue workers, robots, or vehicles. These layers are intended to facilitate the construction of applications for these networks. Our work has been mainly theoretical, but has recently moved in the direction of application to practical communication and control problems.

Virtual Nodes are simple abstract machines that can be emulated by real physical nodes, and that can be programmed easily, to build applications.

4.3.1 Geoquorums

In our first paper using Virtual Nodes, the VNs are simply passive objects at fixed, known locations. We used them to implement reconfigurable atomic memory. This work was presented at DISC03 and subsequently invited to appear in the special journal issue based on that conference. This work also appears in Part I of Seth Gilbert's PhD thesis, discussed in Section 4.4.

Shlomi Dolev, Seth Gilbert, Nancy A. Lynch, Alex A. Shvartsman, and Jennifer L. Welch. GeoQuorums: Implementing Atomic Memory in Mobile Ad Hoc Networks. Distributed Computing, Special Issue DISC03, 18(2):125-155, 2005. Also, Technical Report MIT-LCS-TR-900a, CSAIL, Massachusetts Institute of Technology, Cambridge, MA, 2004.

Abstract: We present a new approach, the GeoQuorums approach, for implementing atomic read/write shared memory in mobile ad hoc networks. Our approach is based on associating abstract atomic objects with certain geographic locations. We assume the existence of focal points, geographic areas that are normally "populated" by mobile nodes. For example, a focal point may be a road junction, a scenic observation point, or a water resource in the desert. Mobile nodes that happen to populate a focal point participate in implementing a shared atomic object, using a replicated state machine approach. These objects, which we call focal point objects, are then used to implement atomic read/write operations on a virtual shared object, using our new GeoQuorums algorithm. The GeoQuorums algorithm uses a quorum-based strategy in which each each quorum consists of a set of focal point objects. The quorums are used to maintain the consistency of the shared memory and to tolerate limited failures of the focal point objects, caused by depopulation of the corresponding geographic areas. We present a mechanism for changing the set of quorums on the fly, thus improving efficiency. Overall, the new GeoQuorums algorithm efficiently implements read and write operations in a highly dynamic, mobile network.

D. Tulone. Is it possible to ensure strong data guarantees in highly mobile networks? In Proc. of the 5th Annual Mediterranean Workshop of Ad hoc Networks (MedHoc), June 2006.

Abstract: Ensuring the consistency and the availability of replicated data in highly mobile ad hoc networks is a challenging task because of the lack of a backbone infrastructure. Previous works provide strong data guarantees by limiting the motion and the speed of the mobile nodes during the entire system lifetime, and by relying on assumptions that are not realistic for most mobile applications. In this paper we provide a small set of mobility constraints necessary to ensure strong data guarantees. Our constraints can be applied also to low density mobile networks and to applications where the speed and the motion of the mobile nodes are unknown and they can change suddenly, such as vehicular networks. Our mobility model allows us to implement a read/write atomic shared memory that is able to guarantee data availability and atomic consistency despite high node mobility and node failures. Our implementation is provably correct and it can be applied for instance to energy management and to task coordination, as we show in the paper.

4.3.2 Virtual Mobile Nodes

Our next contribution to the area was the definition of more powerful Virtual Nodes that are active state machines rather than just passive objects, and that can themselves move, according to preplanned trajectories. These can be used for communication and sensor data collection, as well as intelligent highway applications. This work appeared in DISC04, and in Part I of Seth Gilbert's PhD thesis (which is discussed in Section 4.4).

In addition, we have briefly considered Virtual Mobile Nodes with trajectories that are computed on-the-fly.

Shlomi Dolev, Seth Gilbert, Nancy Lynch, Elad Schiller, Alex Shvartsman, Jennifer Welch. Brief Announcement: Virtual Mobile Nodes for Mobile Ad Hoc Networks. Proceedings of the 23rd Annual ACM SIGACT-SIGOPS Symposium on Principles of Distributed Computing (PODC 2004), St. John's, Newfoundland, Canada, page 385, July 25-28, 2004. Also, Technical Report MIT-LCS-TR-937, MIT CSAIL, Cambridge, MA, 2004.

Shlomi Dolev, Seth Gilbert, Nancy A. Lynch, Elad Schiller, Alex A. Shvartsman, and Jennifer L. Welch. Virtual Mobile Nodes for Mobile Ad Hoc Networks. 18th International Symposium on Distributed Computing (DISC04), Trippenhuis, Amsterdam, the Netherlands, October 4-7, 2004. Also, in Rachid Guerraoui, editor, volume 3274 of Lecture Notes in Computer Science, Springer-Verlag, December 2004.

Abstract. One of the most significant challenges introduced by mobile networks is coping with the unpredictable motion and the unreliable behavior of mobile nodes. In this paper, we define the

Virtual Mobile Node Abstraction, which consists of robust virtual nodes that are both predictable and reliable. We present the Mobile Point Emulator, a new algorithm that implements the Virtual Mobile Node Abstraction. This algorithm replicates each virtual node at a constantly changing set of real nodes, modifying the set of replicas as the real nodes move in and out of the path of the virtual node. We show that the Mobile Point Emulator correctly implements a virtual mobile node, and that it is robust as long as the virtual node travels through well-populated areas of the network. The Virtual Mobile Node Abstraction significantly simplifies the design of efficient algorithms for highly dynamic mobile ad hoc networks.

Shlomi Dolev, Seth Gilbert, Elad Schiller, Alex Shvartsman, Jennifer Welch. Brief Announcement: Autonomous Virtual Mobile Nodes. 17th ACM Symposium on Parallelism in Algorithms and Architectures (SPAA 2005), Las Vegas, Nevada, July 2005.

Shlomi Dolev, Seth Gilbert, Elad Schiller, Alex A. Shvartsman, and Jennifer Welch. Autonomous Virtual Mobile Nodes. DIAL-M-POMC 2005: Third Annual ACM/SIGMOBILE International Workshop on Foundation of Mobile Computing, Cologne, Germany, September 2, 2005.

Shlomi Dolev, Seth Gilbert, Elad Schiller, Alex Shvartsman, and Jennifer Welch. Autonomous Virtual Mobile Nodes. Technical Report MIT-LCS-TR-992, MIT CSAIL, Cambridge, MA, June 2005.

Abstract: This paper presents a new abstraction for virtual infrastructure in mobile ad hoc networks. An Autonomous Virtual Mobile Node (AVMN) is a robust and reliable entity that is designed to cope with the inherent difficulties caused by processors arriving, leaving, and moving according to their own agendas, as well as with failures and energy limitations. There are many types of applications that may make use of the AVMN infrastructure: tracking, supporting mobile users, or searching for energy sources. The AVMN extends the focal point abstraction in DGLSW03 and the virtual mobile node abstraction in DGLSSW04. The new abstraction is that of a virtual general-purpose computing entity, an automaton that can make autonomous on-line decisions concerning its own movement. We describe a self-stabilizing implementation of this new abstraction that is resilient to the chaotic behavior of the physical processors and provides automatic recovery from any corrupted state of the system.

4.3.3 Virtual Stationary Nodes

The most important special case of Virtual Mobile Nodes is the case of Virtual Stationary Nodes—active state machines that reside at fixed, known locations, e.g., at the intersection points of a grid. Our papers on this special case further developed the notion of an active Virtual Node by giving it some extra power over the timing of its own actions, and by considering self-stabilization of its implementation and applications. The study of self-stabilization for Virtual Nodes is still ongoing, and will form the core of Tina Nolte's PhD thesis.

In addition to defining the basic notion of Virtual Stationary Nodes, and designing emulation algorithms, we produced several applications of VSNs to important communication problems such as geocast and point-to-point communication, and to intruder tracking. We also produced an actual implementation of simple Virtual Nodes, running on IPaq hand-held mobile computers. This has evolved into a simulation of VSNs, developed by Mike Spindel for his MEng thesis.

Virtual Nodes appear to provide a practical approach to point-to-point message routing in mobile ad hoc networks. Mike Spindel is using his platform to compare VN-based implementations of mobile ad hoc network message routing with alternative algorithms that have been proposed. Also, this past summer, Calvin Newport worked with Ralph Droms at Cisco to explore the feasibility of implementing actual IPv6 message routing over a Virtual Node substrate; a preliinary publication will be forthcoming.

Shlomi Dolev, Seth Gilbert, Limor Lahiani, Nancy Lynch, and Tina Nolte. Virtual Stationary Automata for Mobile Networks (Extended Abstract) Technical Report MIT-LCS-TR-979, MIT CSAIL, Cambridge, MA 02139, January 2005.

Shlomi Dolev, Limor Lahiani, Seth Gilbert, Nancy Lynch, Tina Nolte. Brief Announcement: Virtual Stationary Automata for Mobile Networks. Proceedings of the 24th Annual ACM Symposium on Principles of Distributed Computing (PODC'05), Las Vegas, Nevada, July, 2005.

The task of designing algorithms for constantly changing networks is difficult. We focus on mobile ad-hoc networks, where mobile processors attempt to coordinate despite min- imal infrastructure support. We develop new techniques to cope with this dynamic, heterogeneous, and chaotic environment. We mask the unpredictable behavior of mobile networks by defining and emulating a virtual infrastructure, consisting of timing-aware and location-aware machines at fixed locations, that mobile nodes can interact with. The static virtual infrastructure allows appplication developers to use simpler algorithms including many previously developed for fixed networks.

Shlomi Dolev, Seth Gilbert, Limor Lahiani, Nancy Lynch, and Tina Nolte. Timed Virtual Stationary Automata for Mobile Networks. Techncial Report MIT-LCS-TR-979a, MIT CSAIL, Cambridge, MA, August 2005.

Shlomi Dolev, Seth Gilbert, Limor Lahiani, Nancy Lynch, and Tina Nolte. Timed Virtual Stationary Automata for Mobile Networks. Allerton Conference 2005: Forty-Third Annual Allerton Conference on Communication, Control, and Computing, September 2005. Invited paper.

Shlomi Dolev, Seth Gilbert, Limor Lahiani, Nancy Lynch, and Tina Nolte. Timed Virtual Stationary Automata for Mobile Networks. 9th International Conference on Principles of Distributed Systems (OPODIS 2005), Pisa, Italy, December 12-14, 2005.

Abstract: We define a programming abstraction for mobile networks called the Timed Virtual Stationary Automata programming layer, consisting of mobile clients, virtual timed I/O automata called virtual stationary automata (VSAs), and a communication service connecting VSAs and client nodes. The VSAs are located at prespecified regions that tile the plane, defining a static virtual infrastructure. We present a self-stabilizing algorithm to emulate a timed VSA using the real mobile nodes that are currently residing in the VSAs region.We also discuss examples of applications whose implementations benefit from the simplicity obtained through use of the VSA abstraction.

Shlomi Dolev, Limor Lahiani, Nancy Lynch, and Tina Nolte. Self-Stabilizing Mobile Node Location Management and Message Routing. SSS 2005: Seventh International Symposium on Self-Stabilizing Systems, Barcelona, Spain, October 2005.

Shlomi Dolev, Limor Lahiani, Nancy Lynch, and Tina Nolte. Self-Stabilizing Mobile Node Location Management and Message Routing. Technical Report MIT-LCS-TR-999, MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, August 2005.

Abstract: We present simple algorithms for achieving self-stabilizing location management and routing in mobile ad-hoc networks. While mo- bile clients may be susceptible to corruption and stopping failures, mobile networks are often deployed with a reliable GPS oracle, supplying frequent updates of accurate real time and location information to mobile nodes. Information from a GPS oracle provides an external, shared source of consistency for mobile nodes, allowing them to label and timestamp messages, and hence aiding in identification of, and eventual recovery from, corruption and failures. Our algorithms use a GPS oracle. Our algorithms also take advantage of the Virtual Stationary Automata programming abstraction, consisting of mobile clients, virtual timed ma- chines called virtual stationary automata (VSAs), and a local broadcast service connecting VSAs and mobile clients. VSAs are distributed at known locations over the plane, and emulated in a self-stabilizing manner by the mobile nodes in the system. They serve as fault-tolerant build- ing blocks that can interact with mobile clients and each other, and can simplify implementations of services in mobile networks. We implement three self-stabilizing, fault-tolerant services, each built on the prior services: (1) VSA-to-VSA geographic routing, (2) mobile client location management, and (3) mobile client end-to-end routing. We use a greedy version of the classical depth-first search algorithm to route messages between VSAs in different regions. The mobile client location management service is based on home locations: Each client identifier hashes to a set of home locations, regions whose VSAs are periodically updated with the clients location. VSAs maintain this information and answer queries for client locations. Finally, the VSA-to-VSA routing and location management services are used to implement mobile client end- to-end routing.

Tina Nolte and Nancy Lynch. A Virtual Node-Based Tracking Algorithm for Mobile Networks. International Conference on Distributed Computing Systems (ICDCS 2007), Toronto, Canada, June, 2007.

Abstract: We introduce a virtual-node based mobile object tracking algorithm for mobile sensor networks, VINESTALK. The algorithm uses the *Virtual Stationary Automata* programming layer, consisting of mobile clients, virtual timed machines distributed at known locations in the plane, called virtual stationary automata (VSAs), and a communication service connecting VSAs and mobile clients.

VINESTALK maintains a data structure on top of an underlying hierarchical partitioning of the network. In a grid partitioning, operations to *find* a mobile object distance d away take $O(d)$ time and communication to complete. Updates to the tracking structure after the object has moved a total of d distance take $O(d * log$network diameter) amortized time and communication to complete. The tracked object may relocate without waiting for VINESTALK to complete updates for prior moves, and while a *find* is in progress.

Tina Nolte and Nancy Lynch. Self-stabilization and Virtual Node Layer Emulations. Proceedings of SSS 2007: Ninth International Symposium on Stabilization, Safety, and Security of Distributed Systems, Paris, France, November 2007.

Abstract: We present formal definitions of stabilization for the Timed I/O Automata (TIOA) framework, and of emulation for the timed Virtual Stationary Automata programming abstraction layer, which consists of mobile clients, virtual timed machines called virtual stationary automata (VSAs), and a local broadcast service connecting VSAs and mobile clients. We then describe what it means for mobile nodes with access to location and clock information to emulate the VSA layer in a self-stabilizing manner. We use these definitions to prove basic results about executions of self-stabilizing algorithms run on self-stabilizing emulations of a VSA layer, and apply these results to a simple geographic routing algorithm running on the VSA layer.

Matthew Brown, Seth Gilbert, Nancy Lynch, Calvin Newport, Tina Nolte, and Michael Spindel. The Virtual Node Layer: A Programming Abstraction for Wireless Sensor Networks. Proceedings of the the International Workshop on Wireless Sensor Network Architecture (WWSNA), Cambridge, MA, April, 2007.

Abstract: The Virtual Node Layer (VNLayer) programming abstraction provides programmable, predictable automata—virtual nodes—emulated by the low-level network nodes. This simplifies the design and rigorous analysis of applications for the wireless sensor network setting, as the layer can mask much of the uncertainty of the underlying components. In this paper, we define a general VNLayer architecture, and then use this framework to design a practical VNLayer implementation, optimized for real-world use. We then discuss our experience deploying this implementation on a testbed of hand-held computers, and in a custom-built packet-level simulator, and present a sample application—a Virtual Traffic Light—to highlight the power and utility of our abstraction. We conclude with a survey of additional applications that are well suited to this setting.

4.3.4 Using Virtual Nodes for motion coordination

An important potential use that has emerged for (stationary) Virtual Nodes is coordinating the motion of robots, vehicles, aircraft, etc. We first produced a paper describing a simple method of coordinating robots to form a pattern. We then developed a more detailed description of a VN-based air-traffic control system.

A VN-based air-traffic control system can be considered a generalization of a Virtual Traffic Light, as described in the previous subsection.

Nancy Lynch, Sayan Mitra, and Tina Nolte. Motion coordination using virtual nodes. Technical Report MIT-LCS-TR-986, MIT CSAIL, Cambridge, MA, April 2005.

Nancy Lynch, Sayan Mitra, and Tina Nolte. Motion Coordination Using Virtual Nodes. CDC-ECC 2005: Forty-Fourth IEEE Conference on Decision and Control and European Control Conference, Seville, Spain, December 2005. Invited paper.

Abstract: We describe how a virtual node abstraction layer can be used to coordinate the motion of real mobile nodes in a region of 2-space. In particular, we consider how nodes in a mobile ad hoc network can arrange themselves along a predetermined curve in the plane, and can maintain themselves in such a configuration in the presence of changes in the underlying mobile ad hoc network, specifically, when nodes may join or leave the system or may fail. Our strategy is to allow the mobile nodes to implement a virtual layer consisting of mobile client nodes, stationary Virtual Nodes (VNs) for predetermined zones in the plane, and local broadcast communication. The VNs coordinate among themselves to distribute the client nodes between zones based on the length of the curve through those zones, while each VN directs its zone's local client nodes to move themselves to equally spaced locations on the local portion of the target curve.

Matthew D. Brown. Air Traffic Control Using Virtual Stationary Automata. Master of Engineering Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, September 2007.

Abstract: As air travel has become an essential part of modern life, the air traffic control system has become strained and overworked. This problem is occurring because the capacity of the current air traffic control system is severely limited by the capabilities of its human operators. Therefore, if we are to increase the capacity of the air traffic control system, then we must develop new automated systems for air traffic control. In my thesis, I take a distributed approach to automated air traffic control. I use a wireless ad-hoc network to simulate a layer of Virtual Stationary Automata, or VSAs, which are virtual machines located at fixed locations in space. These VSAs can then be used to help coordinate the aircraft in the air traffic control system. I model the air traffic control system as a directed graph, showing how the continuous real world air traffic can be abstracted into a discrete graph representation. Using this graph representation, I provide two algorithms to perform safe and efficient air traffic control and prove their effectiveness.

4.4 Reliable Computing in Unreliable Networks

Most theoretical algorithms for mobile networks, including our own, are based on strong assumptions about the MAC communication layer. For example, much of our work on Virtual Nodes assumes totally-ordered, reliable local broadcast. We have begun a study of algorithms and computability for mobile networks with weaker assumptions about the MAC layer, such as message loss and collisions.

Our initial work in this area studied distributed consensus and related problems in single-hop unreliable wireless networks. This work has progressed to the point of developing a complete emulation of a Virtual Node layer in an unreliable multi-hop network. This last piece of work forms Part II of Seth Gilbert's PhD thesis.

We have continued this work with an ongoing study of communication capacity in the presence of Byzantine faulty processes.

Gregory Chockler, Murat Demirbas, Seth Gilbert, Nancy Lynch, Calvin Newport, and Tina Nolte. Reconciling the Theory and Practice of (Un)Reliable Wireless Broadcast. Proceedings of the 4th International Workshop on Assurance in Distributed Systems and Networks (ADSN 2005), June 6, 2005, Columbus, Ohio, USA.

Abstract: Theorists and practitioners have fairly different perspectives on how wireless broadcast works. Theorists think about synchrony; practitioners think about backoff. Theorists assume reliable communication; practitioners worry about collisions. The examples are endless. Our goal is to begin to reconcile the theory and practice of wireless broadcast, in the presence of failures. We propose new models for wireless broadcast and use them to examine what makes a broadcast model good. In the process, we pose some interesting questions that will help to bridge the gap.

G. Chockler, M. Demirbas, S. Gilbert, and C. Newport. A Middleware Framework for Robust Applications in Wireless Ad Hoc Networks. Allerton Conference 2005: Forty-Third Annual Allerton Conference on Communication, Control, and Computing, September 2005.

Abstract: Wireless ad hoc networks are becoming an increasingly common platform for bringing computation to environments with minimal infrastructure. Increasingly, applications require robust fault-tolerance guarantees, despite a challenging network environment. In this paper, we introduce a new middleware framework for wireless ad hoc networks to aid the development of robust algorithms. Our framework is based on the following three components: (1) receiver-side collision detection, used for identifying inconsistencies caused by unreliable communication; (2) robust round synchronization, used for emulating a strictly synchronized multi-hop network using only basic timeliness assumptions about the environment; and (3) contention management, used for reducing message collision and supporting eventually reliable message delivery. We demonstrate the utility of our framework by showing how it can be used to implement a simple fault-tolerant broadcast protocol, and discuss algorithms to implement each of the components.

Gregory Chockler, Murat Demirbas, Seth Gilbert, Calvin Newport, and Tina Nolte. Consensus and Collision Detectors in Wireless Ad Hoc Networks. Twenty-Fourth Annual ACM SIGACT-SIGOPS Symposium on Principles of Distributed Computing (PODC 2005), Las Vegas, Nevada, pages 197-206, July 2005.

Abstract: We consider the fault-tolerant consensus problem in wireless ad hoc networks with crash-prone nodes. We develop consensus algorithms for single-hop environments where the nodes are located within broadcast range of each other. Our algorithms tolerate highly unpredictable wireless communication, in which messages may be lost due to collisions, electromagnetic interference, or other anomalies. Accordingly, each node may receive a different set of messages in the same round. In order to minimize collisions, we design adaptive algorithms that attempt to minimize the broadcast contention. To cope with unreliable communication, we augment the nodes with collision detectors and present a new classification of collision detectors in terms of accuracy and completeness, based on practical realities. We show exactly in which cases consensus can be solved, and thus determine the requirements for a useful collision detector. We validate the feasibility of our algorithms, and the underlying wireless model, with simulations based on a realistic 802.11 MAC layer implementation and a detailed radio propagation model. We analyze the performance of our algorithms under varying sizes and densities of deployment and varying MAC layer parameters. We use our single-hop consensus algorithms as the basis for solving consensus in a multi-hop network, demonstrating the resilience of our algorithms to a challenging and noisy environment.

Calvin Newport. Consensus and Collision Detectors in Wireless Ad Hoc Networks. Masters Thesis, MIT Department of Electrical Engineering and Computer Science, Cambridge, MA, June 2006.

Abstract: In this study, we consider the fault-tolerant consensus problem in wireless ad hoc networks with crashprone nodes. Specifically, we develop lower bounds and matching upper bounds for this problem in single-hop wireless networks, where all nodes are located within broadcast range of each other. In a novel break from existing work, we introduce a highly unpredictable communication model in which each node may lose an arbitrary subset of the messages sent by its neighbors during each round. We argue that this model better matches behavior observed in empirical studies of these networks. To cope with this communication unreliability we augment nodes with receiver-side collision detectors and present a new classification of these detectors in terms of accuracy and completeness. This classification is motivated by practical realities and allows us to determine, roughly speaking, how much collision detection capability is enough to solve the consensus problem efficiently in this setting. We consider ten different combinations of completeness and accuracy properties in total, determining for each whether consensus is solvable, and, if it is, a lower bound on the number of rounds required. Furthermore, we distinguish anonymous and non-anonymous protocolswhere "anonymous" implies that devices do not have unique identifiersdetermining what effect (if any) this extra information has on the complexity of the problem. In all relevant cases, we provide matching upper bounds. Our contention is that the introduction of (possibly weak) receiver-side collision detection is an important approach to reliably solving problems in unreliable networks. Our results, derived in a realistic network model, provide important feedback to ad hoc network practitioners regarding what hardware (and low-layer software) collision detection capability is sufficient to facilitate the construction of reliable and fault-tolerant agreement protocols for use in real-world deployments.

Seth Gilbert. Virtual Infrastructure for Wireless Ad Hoc Networks. PhD Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, 2007.

Abstract: One of the most significant challenges introduced by ad hoc networks is coping with the unpredictable deployment, uncertain reliability, and erratic communication exhibited by emerging wireless networks and devices. The goal of this thesis is to develop a set of algorithms that address these challenges and simplify the design of algorithms for ad hoc networks.

In the first part of this thesis, I introduce the idea of virtual infrastructure, an abstraction that provides reliable and predictable components in an unreliable and unpredictable environment. This part assumes reliable communication, focusing primarily on the problems created by unpredictable motion and fault-prone devices. I introduce several types of virtual infrastructure, and present new algorithms based on the replicated-state-machine paradigm to implement these infrastructural components.

In the second part of this thesis, I focus on the problem of developing virtual infrastructure for more realistic networks, in particular coping with the problem of unreliable communication. I introduce a new framework for modeling wireless networks based on the ability to detect collisions. I then present a new algorithm for implementing replicated state machines in wireless networks, and show how to use replicated state machines to implement virtual infrastructure even in an environment with unreliable communication.

Michael A. Bender, Jeremy T. Fineman, and Seth Gilbert. Contention Resolution with Heterogeneous Job Sizes. Proceedings of the 14th Annual European Symposium on Algorithms, September, 2006.

Abstract: We study the problem of contention-resolution for different-sized jobs on a simple channel. When a job makes a run attempt on a simple channel, it learns only whether the attempt succeeds or fails. We first analyze the binary exponential backoff protocol, and show that it achieves a makespan of $V2^{\Theta}(\sqrt{logn})$, where V is the total work of all the contending jobs. This bound is significantly larger than when all jobs are constant-sized. We then analyze a variant of exponential backoff that achieves makespan $O(VlogV)$. Finally, we introduce a new protocol, called size-hashed backoff, specifically designed for jobs of multiple sizes that achieves makespan $O(Vlog^3logV)$.

Seth Gilbert, Rachid Guerraoui, and Calvin Newport. Of Malicious Motes and Suspicious Sensors. Technical Report MIT-CSAIL-TR-2006-026, CSAIL, Massachusetts Institute of Technology, Cambridge, MA, April 2006.

Seth Gilbert, Rachid Guerraoui, and Calvin Newport. Of Malicious Motes and Suspicious Sensors: On the Efficiency of Malicious Interference in Wireless Networks. Proceedings of the 10th International Conference On Principles Of Distributed Systems (OPODIS), Bordeaux, France, December 12-15, 2006.

Abstract: How efficiently can a malicious device disrupt a single-hop wireless network? Imagine a game involving two honest players, Alice and Bob, who want to exchange information, as well as a malicious adversary, Collin, who wants to prevent them from communicating. Previous work assumes that the adversary cannot induce collisions in the network. By contrast, we allow Collin a budget of beta broadcasts, which he can use to arbitrarily disrupt communication. We show that Alice and Bob can be delayed for exactly $2beta + Theta(lg|V|)$ communication rounds, where V is the set of values that Alice and Bob may transmit. From this we derive bounds on Collin's efficiency, showing an inherent "jamming gain" of 2, and "disruption-free complexity" of $Theta(lg|V|)$. The trials and tribulations of Alice and Bob in fact capture something fundamental about how efficiently malicious devices can disrupt wireless communication. We derivevia reduction to the 3-player gameround complexity lower bounds for several classical n-player problems: $2beta + Theta(lg|V|)$ for reliable broadcast, $2beta + Omega(logn/k)$ for leader election among k contenders, and $2beta + Omega(klg|V|/k)$ for static k-selection. Then, we consider an extension of our adversary model that also includes up to t crash failures. We study binary consensus as the archetypal problem for this environment and show a bound of $2beta + Theta(t)$ rounds. These results imply immediate bounds on jamming gain and disruption-free complexity. We provide tight, or nearly tight, upper bounds for all four problems.

Shlomi Dolev, Seth Gilbert, Rachid Guerraoui, and Calvin Newport. Gossiping in a Multi-Channel Radio Network: An Oblivious Approach to Coping with Malicious Interference. Proceedings of the 21th International Symposium on Distributed Computing (DISC), Lemesos, Cyprus, September, 2007.

Abstract: We study oblivious deterministic gossip algorithms for multi-channel radio networks with a malicious adversary. In a multi-channel network, each of the n processes in the system must choose, in each round, one of the c channels of the system on which to participate. Assuming the adversary can disrupt one channel per round, preventing communication on that channel, we establish a tight bound on the number of rounds needed to solve the $\epsilon - gossip$ problem , a parameterized generalization of the all-to-all gossip problem that requires $(1-\epsilon)n$ of the "rumors" to be successfully disseminated. Underlying our lower bound proof lies an interesting connection between $\epsilon - gossip$ and extremal graph theory. Specifically, we make use of Turan's theorem, a seminal result in extremal combinatorics, to reason about an adversary's optimal strategy for disrupting an algorithm of a given duration. We then show how to generalize our upper bound to cope with an adversary that can simultaneously disrupt t channels. Our generalization makes use of selectors: a combinatorial tool that guarantees that any subset of processes will be "selected" by some set in the selector. We prove this generalized algorithm optimal if a maximum number of values is to be gossiped. We conclude by extending our algorithm to tolerate traditional Byzantine corruption faults.

4.5 Other Algorithmic Work

Members of the group also produced a potpourri of other papers on algorithms for mobile and traditional communication networks.

4.5.1 Shared-memory algorithms

Gregory Chockler, Idit Keidar and Dahlia Malkhi. Optimal Resilience Wait-Free Storage from Byzantine Components: Inherent Costs and Solutions. FuDiCo II: S.O.S. Survivability: Obstacles and Solutions. 2nd Bertinoro Workshop on Future Directions in Distributed Computing, 23-25 June 2004 University of Bologna Residential Center Bertinoro (Forli), Italy.

Abstract: This position paper describes the results of our on-going investigation into the possibility and cost of building a survivable store. It considers optimal resilience systems comprising of $3t+1$ base storage units, t of which may fail by becoming non-responsive or arbitrarily corrupted. Our contribution includes both algorithms and lower bounds in this model. We illuminate an inherent difficulty of achieving optimal resilience in the form of two lower bounds, on read and on write complexities. We also provide the first optimal-resilience wait-free algorithms that match these bounds in performance. Finally, we suggest some directions for future research.

Ittai Abraham, Gregory Chockler, Idit Keidar and Dahlia Malkhi. Byzantine Disk Paxos: Optimal Resilience with Byzantine Shared Memory. Proceedings of the 23rd Annual ACM SIGACT-SIGOPS Symposium on Principles of Distributed Computing (PODC 2004), St. John's, Newfoundland, Canada, pages 226-235, July 2004.

Abstract: We present Byzantine Disk Paxos, an asynchronous shared- memory consensus protocol that uses a collection of $n > 3t$ disks, t of which may fail by becoming non-responsive or arbitrarily corrupted. We give two constructions of this protocol; that is, we construct two different building blocks, each of which can be used, along with a leader oracle, to solve consensus. One building block is a shared wait-free safe register. The second building block is a regular register that satisfies a weaker termination (liveness) condition than wait freedom: its write operations are wait-free, whereas its read operations are guaranteed to return only in executions with a finite number of writes. We call this termination condition finite writes (FW), and show that consensus is solvable with FW-terminating registers and a leader oracle. We construct each of these reliable registers from $n > 3t$ base registers, t of which can be non-responsive or Byzantine. All the previous wait-free constructions in this model used at least $4t + 1$ fault-prone registers, and we are not familiar with any prior FW-terminating constructions in this model.

Gregory Chockler and Dahlia Malkhi. Active Disk Paxos with infinitely many processes. Distributed Computing, volume 18, number 1, pages 73-84, July 2005.

Abstract: We present an improvement to the Disk Paxos protocol by Gafni and Lamport which utilizes extended functionality and flexibility provided by Active Disks and supports unmediated concurrent data access by an unlimited number of processes. The solution facilitates coordination by an infinite number of clients using finite shared memory. It is based on a collection of read-modify-write objects with faults, that emulate a new, reliable shared memory abstraction called a ranked register. The required read-modify-write objects are readily available in Active Disks and in Object Storage Device controllers, making our solution suitable for state-of-the-art Storage Area Network (SAN) environments.

4.5.2 Traditional networks

Carl Livadas and Idit Keidar. Caching-Enhanced Scalable Reliable Multicast. International Conference on Dependable Systems and Networks (DSN), June-July 2004.

Abstract: We present the Caching-Enhanced Scalable Reliable Multicast (CESRM) protocol. CESRM augments the Scalable Reliable Multicast (SRM) protocol with a caching-based expedited recovery scheme. CESRM exploits the packet loss locality occurring in IP multicast transmissions in order to expeditiously recover from losses in the manner in which recent losses were recovered. Trace-driven simulations show that CESRM reduces the average recovery latency of SRM by roughly 50in terms of recovery traffic and control messages.

Seth Gilbert and Gregory Malewicz. The Quorum Deployment Problem. OPODIS 2004: 8th International Conference on Principles of Distributed Systems, Grenoble, France, December 15-17, 2004. Also, full version in MIT CSAIL Technical Report MIT-LCS-TR-972, October 2004.

Abstract: Quorum systems are commonly used to maintain the consistency of replicated data in a distributed system. Much research has been devoted to developing quorum systems with good theoretical properties, such as fault tolerance and high availability. However, even given a theoretically good quorum system, it is not obvious how to efficiently deploy such a system in a real network. This paper introduces a new combinatorial optimization problem, the Quorum Deployment Problem, and studies its complexity. We demonstrate that it is NP-hard to approximate the Quorum Deployment Problem within any factor of n^δ, where n is the number of nodes in the distributed network and $\delta > 0$. The problem is NP-hard in even the simplest possible distributed network: a one-dimensional line with metric cost. We begin to study algorithms for variants of the problem. Some variants can be solved optimally in polynomial time and some NP-hard variants can be approximated to within a constant factor.

Gregory Chockler, Seth Gilbert, and Boaz Patt-Shamir. Communication-Efficient Probabilistic Quorum Systems. Proceedings of the IEEE International Workshop on Foundations and Algorithms for Wireless Networking (FAWN), March 29-31, 2006.

Abstract: Communication-efficiency is of key importance when constructing robust services in limited bandwidth environments, such as sensor networks. We focus on communication-efficiency in the context of quorum systems, which are useful primitives for building reliable distributed systems. To this end, we exhibit a new probabilistic quorum construction in which every node transmits at most $O(log2n)$ bits per quorum access, where n is the number of nodes in the system. Our implementation, in addition to being communication efficient, is also robust in the face of communication failures. In particular, it guarantees consistency (with high probability) in the face of network partitions. To the best of our knowledge, no existing probabilistic quorum systems achieve polylogarithmic communication complexity and are resilient to network partitions.

A. Russell and A. Shvartsman. Distributed Computation Meets Design Theory: Local Scheduling for Disconnected Cooperation. Current Trends in Theoretical Computer Science: The Challenge of the New Century, vol. 1: Algorithms and Complexity, pp. 315-336, World Scientific, 2004.

Abstract: Ability to cooperate on common tasks in a distributed setting is key to solving a broad range of computation problems ranging from distributed search such as SETI to distributed simulation and multi-agent collaboration. In such settings there exists a trade-off between computation and communication: both resources must be managed to decrease redundant computation and to ensure efficient computational progress. This survey deals with scheduling issues for distributed collaboration. Specifically, we examine the extreme situation of collaboration without communication. That is, we consider the extent to which efficient collaboration is possible if all resources are directed to computation at the expense of communication. Of course there are also cases where such an extreme situation is not a matter of choice: the network may fail, the mobile nodes may have intermittent connectivity, and when communication is unavailable it may take a long time to (re)establish connectivity. The results summarized here precisely characterize the ability of distributed agents to collaborate on a known collection of independent tasks by means of local scheduling decisions that require no communication and that achieve low redundancy in task executions. Such scheduling solutions exhibit an interesting connection between the distributed collaboration problem and the mathematical design theory. The lower bounds presented here along with the randomized and deterministic schedule constructions show the limitations on such low-redundancy cooperation and show that schedules with near-optimal redundancy can be efficiently constructed by processors working in isolation. We also show that when processors start working in isolation and are subjected to an arbitrary pattern of network reconfigurations, e.g., fragmentations and merges, randomized scheduling is competitive compared to an optimal algorithm that is aware of the pattern of reconfigurations.

4.6 Traditional distributed computing theory

While working on dynamic algorithms, we have continued research involving fundamental open problems in (fixed-network) distributed computing theory. Our greatest success in this direction has been the discovery of a new lower bound on the time to reach mutual exclusion. This paper won the Best Student Paper award at PODC 06. A second success has been a new, very simple impossibility proof for the problem of boosting resiliency of distributed services. Finally, we have devised a new failure detector service that is, in a certain sense, weakest for solving the problem of k-consensus.

Rui Fan and Nancy Lynch. An $\Omega(n \log n)$ Lower Bound on the Cost of Mutual Exclusion. Proceedings of the Twenty-Fifth Annual Symposium on Principles of Distributed Computing (PODC'06), Denver, Colorado, July 2006. Best Student Paper Award.

Abstract: We prove an $(nlogn)$ lower bound on the number of nonbusywaiting memory accesses by any deterministic algorithm solving n process mutual exclusion that communicates via shared registers. The cost of the algorithm is measured in the state change cost model, a variation of the cache coherent model. Our bound is tight in this model. We introduce a novel information theoretic proof technique: We establish a lower bound on the information needed by processes to solve mutual exclusion. Then we relate the amount of information processes can acquire through shared memory accesses to the cost they incur. We believe our proof technique is extensible and intuitive, and may be applied to a variety of other problems and system models.

Paul Attie, Rachid Guerraoui, Petr Kouznetsov, Nancy Lynch, and Sergio Rajsbaum. Impossibility of boosting distributed service resilience. Technical Report MIT-LCS-TR-982, Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, February 2005.

Paul Attie, Rachid Guerraoui, Petr Kouznetsov, Nancy Lynch, and Sergio Rajsbaum. The Impossibility of Boosting Distributed Service Resilience. 25th IEEE International Conference on Distributed Computing Systems (ICDCS 2005), Columbus, OH, pages 39-48, June 6-10, 2005.

Abstract: We prove two theorems saying that no distributed system in which processes coordinate using reliable registers and f-resilient services can solve the consensus problem in the presence of $f+1$ undetectable process stopping failures. (A service is f-resilient if it is guaranteed to operate as long as no more than f of the processes connected to it fail.) Our first theorem assumes that the given services are atomic objects, and allows any connection pattern between processes and services. In contrast, we show that it is possible to boost the resilience of systems solving problems easier than consensus: the kset consensus problem is solvable for $2k-1$ failures using 1-resilient consensus services. The first theorem and its proof generalize to the larger class of failure-oblivious services. Our second theorem allows the system to contain failure-aware services, such as failure detectors, in addition to failure-oblivious services; however, it requires that each failure-aware service be connected to all processes. Thus, $f+1$ process failures overall can disable all the failure-aware services. In contrast, it is possible to boost the resilience of a system solving consensus if arbitrary patterns of connectivity are allowed between processes and failure-aware services: consensus is solvable for any number of failures using only 1-resilient 2-process perfect failure detectors.

Rachid Guerraoui, Maurice Herlihy, Petr Kouznetsov, Nancy Lynch, and Calvin Newport. On the weakest failure detector ever. Proceedings of the Twenty-Sixth Annual ACM Symposium on the Principles of Distributed Computing (PODC), Portland, Oregon, August 2007.

Abstract: Many problems in distributed computing are impossible when no information about process failures is available. It is common to ask what information about failures is necessary and sufficient to circumvent some specific impossibility, e.g., consensus, atomic commit, mutual exclusion, etc. This paper asks what information about failures is needed to circumvent any impossibility and sufficient to circumvent some impossibility. In other words, what is the minimal yet non-trivial failure information. We present an abstraction, denoted , that provides very little failure information. In every run of the distributed system, eventually informs the processes that some set of processes in the system cannot be the set of correct processes in that run. Although seemingly weak, for it might provide random information for an arbitrarily long period of time, and it only excludes one possibility of correct set among many, still captures non-trivial failure information. We show that is sufficient to circumvent the fundamental wait-free set-agreement impossibility. While doing so, we (a) disprove previous conjectures about the weakest failure detector to solve set-agreement and we (b) prove that solving set-agreement with registers is strictly weaker than solving $n + 1$-process consensus using n-process consensus. We prove that is, in a precise sense, minimal to circumvent any wait-free impossibility. Roughly, we show that is the weakest eventually stable failure detector to circumvent any wait-free impossibility. Our results are generalized through an abstraction f that we introduce and prove necessary to solve any problem that cannot be solved in an f-resilient manner, and yet sufficient to solve f-resilient f-set-agreement.

4.7 Mathematical Foundations: Modeling and Verification

Our research on algorithms for dynamic networks has required mathematical foundations, in the form of Timed, Hybrid, and Probabilistic I/O Automata modeling frameworks. For example, modeling the motion of physical components in a mobile network requires a timed or hybrid modeling framework. Modeling safety requirements for physical safety-critical systems requires a framework that can express both hybrid and probabilistic aspects of system behavior.

4.7.1 Timed I/O Automata

In the early part of our AFOSR project, we completed a monograph on the basics of the Timed I/O Automata math framework.

Sayan Mitra and Nancy Lynch. Proving approximate implementation relations for Probabilistic I/O Automata. Electronic Notes in Theoretical Computer Science, 174(8):71-93, 2007.

Abstract: This monograph presents the Timed Input/Output Automaton (TIOA) modeling framework, a basic mathematical framework to support description and analysis of timed (computing) systems. Timed systems are systems in which desirable correctness or performance properties of the system depend on the timing of events, not just on the order of their occurrence. Timed systems are employed in a wide range of domains including communications, embedded systems, real-time operating systems, and automated control. Many applications involving timed systems have strong safety, reliability and predictability requirements, which makes it important to have methods for systematic design of systems and rigorous analysis of timing-dependent behavior. An important feature of the TIOA framework is its support for decomposing timed system descriptions. In particular, the framework includes a notion of external behavior for a timed I/O automaton, which captures its discrete interactions with its environment. The framework also defines what it means for one TIOA to implement another, based on an inclusion relationship between their external behavior sets, and defines notions of simulations, which provide sufficient conditions for demonstrating implementation relationships. The framework includes a composition operation for TIOAs, which respects external behavior, and a notion of receptiveness, which implies that a TIOA does not block the passage of time.

4.7.2 Probabilistic I/O Automata

We continued our work on developing probabilistic I/O automata, emphasizing compositional aspects. We also carried out a smaller effort in defining approximate simulation relations between probabilistic I/O automata.

Nancy Lynch, Roberto Segala, and Frits Vaandrager. Compositionality for Probabilistic Automata. Technical Report MIT-LCS-TR-907, Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, November 2004.

Abstract: We establish that on the domain of probabilistic automata, the trace distribution coincides with the simulation preorder.

Nancy Lynch, Roberto Segala, and Frits Vaandrager. Observing Branching Structure through Probabilistic Contexts. Siam Journal on Computing, 37(4):977-1013, September 2007.

Abstract: Probabilistic automata (PAs) constitute a general framework for modeling and analyzing discrete event systems that exhibit both nondeterministic and probabilistic behavior, such as distributed algorithms and network protocols. The behavior of PAs is commonly defined using schedulers (also called adversaries or strategies), which resolve all nondeterministic choices based on past history. From the resulting purely probabilistic structures, trace distributions can be extracted, whose intent is to capture the observable behavior of a PA. However, when PAs are composed via an (asynchronous) parallel composition operator, a global scheduler may establish strong correlations between the behavior of system components and, for example, resolve nondeterministic choices in one PA based on the outcome of probabilistic choices in the other. It is well known that, as a result of this, the (linear-time) trace distribution precongruence is not compositional for PAs. In his PhD thesis from '95, Segala has shown that the (branching-time) probabilistic simulation preorder is compositional for PAs. In this paper, we establish that the simulation preorder is in fact the coarsest refinement of the trace distribution preorder that is compositional.

We prove our characterization result by providing (1) a context of a given PA A, called the tester, that may announce the state of A to the outside world, and (2) a specific global scheduler, called the observer, which ensures that the state information that is announced is actually correct. Now when another PA B is composed with the tester, it may generate the same external behavior as the observer only when it is able to simulate A in the sense that whenever A goes to some state s, B can go to a corresponding state u from which it may generate the same external behavior. Our result shows that probabilistic contexts together with global schedulers are able to exhibit the branching structure of PAs.

Ling Cheung, Nancy Lynch, Roberto Segala, and Frits Vaandrager. Switched Probabilistic I/O Automata. Nijmegen Institute for Computing and Information Sciences (NIII) Technical Report NIII-R0437, Catholic University of Nijmegen, Nijmegen, The Netherlands, September 2004.

Ling Cheung, Nancy Lynch, Roberto Segala, and Frits Vaandrager. Switched Probabilistic I/O Automata. In Z. Liu and K. Araki, editors, Proceedings of the First International Colloquium on Theoretical Aspects of Computing (ICTAC2004), Guiyang, China, September 2004, volume 3407 of Lecture Notes in Computer Science, pages 494-510, Springer-Verlag, 2005.

Abstract. A switched probabilistic I/O automaton is a special kind of probabilistic I/O automaton (PIOA), enriched with an explicit mechanism to exchange control with its environment. Every closed system of switched automata satisfies the key property that, in any reachable state, at most one component automaton is active. We define a tracebased semantics for switched PIOAs and prove it is compositional. We also propose switch extensions of an arbitrary PIOA and use these extensions to define a new trace-based semantics for PIOAs.

Ling Cheung, Nancy Lynch, Roberto Segala, and Frits Vaandrager. Switched Probabilistic PIOA: Parallel Composition via Distributed Scheduling. Theoretical Computer Science, volume 365, issues 1-2, pages 83-108, 10 November 2006.

Abstract: This paper presents the framework of switched probabilistic input/output automata (or switched PIOA), augmenting the original PIOA framework with an explicit control exchange mechanism. Using this mechanism, we model a network of processes passing a single token among them, so that the location of this token determines which process is scheduled to make the next move. This token structure therefore implements a distributed scheduling scheme: scheduling decisions are always made by the (unique) active component. Distributed scheduling allows us to draw a clear line between local and global nondeterministic choices. We then require that local nondeterministic choices are resolved using strictly local information. This eliminates unrealistic schedules that arise under the more common centralized scheduling scheme. As a result, we are able to prove that our trace-style semantics is compositional.

Sayan Mitra and Nancy Lynch. Approximate simulations for task-structured probabilistic I/O automata. LICS workshop on Probabilistic Automata and Logics (PAul06), Seattle, WA, August 2006.

Abstract: A Probabilistic I/O Automaton (PIOA) is a countable-state automaton model that allows nondeterministic and probabilistic choices in state transitions. A task-PIOA adds a task structure on the locally controlled actions of a PIOA as a means for restricting the nondeterminism in the model. The task-PIOA framework defines exact implementation relations based on inclusion of sets of trace distributions. In this paper we develop the theory of approximate implementations and equivalences for task-PIOAs. We propose a new kind of approximate simulation between task-PIOAs and prove that it is sound with respect to approximate implementations. Our notion of similarity of traces is based on a metric on trace distributions and therefore, we do not require the state spaces nor the space of external actions (output alphabet) of the underlying automata to be metric spaces. We discuss applications of approximate implementations to probabilistic safety verification.

Sayan Mitra and Nancy Lynch. Proving approximate implementation relations for Probabilistic I/O Automata. Electronic Notes in Theoretical Computer Science, 174(8):71-93, 2007.

Abstract: In this paper we introduce the notion of approximate implementations for Probabilistic I/O Automata (PIOA) and develop methods for proving such relationships. We employ a task structure on the locally controlled actions and a task scheduler to resolve nondeterminism. The interaction between a scheduler and an automaton gives rise to a trace distributiona probability distribution over the set of traces. We define a PIOA to be a (discounted) approximate implementation of another PIOA if the set of trace distributions produced by the first is close to that of the latter, where closeness is measured by the (resp. discounted) uniform metric over trace distributions. We propose simulation functions for proving approximate implementations corresponding to each of the above types of approximate implementation relations. Since our notion of similarity of traces is based on a metric on trace distributions, we do not require the state spaces nor the space of external actions of the automata to be metric spaces. We discuss applications of approximate implementations to verification of probabilistic safety and termination.

4.7.3 Combining timed and probabilistic models

A core topic of Sayan Mitra's PhD thesis was the development of mathematical modeling frameworks that support a combination of timed (or hybrid) and probabilistic behavior.

Sayan Mitra and Nancy Lynch. Trace-based semantics of Probabilistic timed I/O automata. Hybrid Systems: Computation and Control (HSCC 2007), Pisa, Italy, April 3-5, 2007, volume 4416 of Lecture Notes in Computer Science. Springer, 2007.

Abstract. We describe the main features of the Probabilistic Timed I/O Automata (PTIOA) a framework for modeling and analyzing discretely communicating probabilistic hybrid systems. A PTIOA can choose the post-state of a discrete transition either nondeterministically or according to (possibly continuous) probability distributions. The framework supports modeling of large systems as compositions of concurrently executing PTIOAs, which interact through shared transition labels. We develop a trace-based semantics for PTIOAs and show that PTIOAs are compositional with respect a new notion of external behavior.

Sayan Mitra. A Verification Framework for Ordinary and Probabilistic Hybrid Systems. PhD Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, September 2007.

Abstract: Combining discrete state transitions with differential equations, Hybrid system models provide an expressive formalism for describing software systems that interact with a physical environment. Automatically checking properties, such as invariance and stability, is extremely hard for general hybrid models, and therefore current research focuses on models with restricted expressive power. In this thesis we take a complementary approach by developing proof techniques that are not necessarily automatic, but are applicable to a general class of hybrid systems. Three components of this thesis, namely, (i) semantics for ordinary and probabilistic hybrid models, (ii) methods for proving invariance, stability, and abstraction, and (iii) software tools supporting (i) and (ii), are integrated within a common mathematical framework.

(i) For specifying nonprobabilistic hybrid models, we present Structured Hybrid I/O Automata (SHIOAs) which adds control theory-inspired structures, namely state models, to the existing Hybrid I/O Automata, thereby facilitating description of continuous behavior. We introduce a generalization of SHIOAs which allows both nondeterministic and stochastic transitions and develop the trace-based semantics for this framework. (ii) We present two techniques for establishing lower-bounds on average dwell time (ADT) for SHIOA models. This provides a sufficient condition of establishing stability for SHIOAs with stable state models. A new simulation-based technique which is sound for proving ADT-equivalence of SHIOAs is proposed. We develop notions of approximate implementation and corresponding proof techniques for Probabilistic I/O Automata. Specifically, a PIOA A is an approximate implementation of B, if every trace distribution of A is close to some trace distribution of B, closeness being measured by a metric on the space of trace distributions. We present a new class of real-valued simulation functions for proving approximate implementations, and demonstrate their utility in quantitatively reasoning about probabilistic safety and termination. (iii) We introduce a specification language for SHIOAs and a theorem prover interface for this language. The latter consists of a translator to typed high order logic and a set of PVS-strategies that partially automate the above verification techniques within the PVS theorem prover.

4.7.4 Stability of hybrid systems

Sayan Mitra worked with Liberzon of U. Illinois to formulate and extend certain stability analysis techniques for hybrid systems using Hybrid I/O Automata.

Sayan Mitra and Daniel Liberzon. Stability of Hybrid Automata with Average Dwell Time: An Invariant Approach. Proceedings of the 43rd Conference on Decision and Control, Paradise Island, Bahamas, December, 2004.

Abstract: A formal method based technique is presented for proving the average dwell time property of a hybrid system, which is useful for establishing stability under slow switching. The Hybrid Input/Output Automaton (HIOA) of [LSV] is used as the model for hybrid systems, and it is shown that some known stability theorems from system theory can be adapted to be applied in this framework. The average dwell time property of a given automaton is formalized as an invariant of a corresponding transformed automaton, such that the former has average dwell time if and only if the latter satisfies the invariant. Formal verification techniques can be used to check this invariance property. In particular, the HIOA framework facilitates inductive invariant proofs by systematically breaking them down into cases for the discrete actions and continuous trajectories of the automaton. The invariant approach to proving the average dwell time property is illustrated by analyzing the hysteresis switching logic unit of a supervisory control system.

Sayan Mitra, Daniel Liberzon, and Nancy Lynch. Verifying average dwell time by solving optimization problems. In Ashish Tiwari and Joao P. Hespanha, editors, Hybrid Systems: Computation and Control (HSCC 06) Santa Barbara, CA, March 2006, volume 3927 of Lecture Notes in Computer Science, pages 476-490, 2006. Springer.

Abstract. In the switched system model, discrete mechanisms of a hybrid system are abstracted away in terms of an exogenous switching signal which brings about the mode switches. The Average Dwell time (ADT) property denes restricted classes of switching signals which provide sufficient conditions for proving stability of switched systems. In this paper, we use a specialization of the Hybrid I/O Automaton model to capture both the discrete and the continuous mechanisms of hybrid systems. Based on this model, we develop methods for automatically verifying ADT properties and present simulation relations for establishing equivalence of hybrid systems with respect to ADT. Given a candidate ADT for a hybrid system, we formulate an optimization problem; a solution of this problem either establishes the ADT property or gives an execution fragment of the system that violates it. For two special classes of hybrid systems, we show that the corresponding optimization problems can be solved using standard mathematical programming techniques.We formally dene equivalence of two hybrid systems with respect to ADT and present a simulation relation-based method for proving this equivalence. The proposed methods are applied to verify ADT properties of a linear hysteresis switch and a nondeterministic thermostat.

4.7.5 Applications and case studies

As we have developed our formal methods, we have tested their usefulness by applying them to a variety of real applications. For example, we have tested our assertional and abstraction methods for (timed and untimed) I/O automata by using them to verify typical algorithms for implementing atomic objects in shared-memory systems. Also, working with network testing expert Nancy Griffeth, we have applied similar methods to verify certain important practical communication protocol designs. Also, working with Ralph Droms of Cisco, we have carried out a major project to abstract and validate a practical fault-tolerant version of the DHCP IP-address management protocol. Working with personnel at NASA, we have used our abstraction and timing methods to validate the SATS small-aircraft landing protocol. Finally, we have utilized new approximate implementation techniques of Mitra, described above, to verify statistical zero-knowledge properties of security protocols.

Gregory Chockler, Nancy Lynch, Sayan Mitra, and Joshua Tauber. Proving Atomicity: An Assertional Approach. Technical Report MIT-CSAIL-TR-2005-048 (and MIT-LCS-TR-995), MIT CSAIL, Cambridge, MA, July 2005.

Gregory Chockler, Nancy Lynch, Sayan Mitra, and Joshua Tauber. Proving Atomicity: An Assertional Approach DISC 2005: 19th International Symposium on Distributed Computing, Cracow, Poland, September 2005.

Abstract: Atomicity (or linearizability) is a commonly used consistency criterion for distributed services and objects. Although atomic object implementations are abundant, proving that algorithms achieve atomicity has turned out to be a challenging problem. In this paper, we initiate the study of systematic ways of verifying distributed implementations of atomic objects, beginning with read/write objects (registers). Our general approach is to replace the existing operational reasoning about events and partial orders with assertional reasoning about invariants and simulation relations. To this end, we define an abstract state machine that captures the atomicity property and prove correctness of the object implementations by establishing a simulation mapping between the implementation and the specification automata. We demonstrate the generality of our specification by showing that it is implemented by three different read/write register constructions: the message-passing register emulation of Attiya, Bar-Noy and Dolev, its optimized version based on real time, and the shared memory register construction of Vitanyi and Awerbuch. In addition, we show that a simplified version of our specification is implemented by a general atomic object construction based on the Lamport's replicated state machine algorithm.

Constantinos Djouvas, Nancy D. Griffeth, and Nancy A. Lynch. Testing Self-Similar Networks MBT 2006: Second Workshop on Model Based Testing, Electronic Notes in Theoretical Computer Science, issue 4, volume 164, March 2006.

Abstract: Network testing presents different challenges from software testing. One challenge is that only a small number of networks, at best, can actually be tested, even when the goal is to test a class of networks. One solution is to select a representative network, which will display any faults present in any network of the class. This paper introduces the use of "self-similarity" to select such a network.

Rui Fan, Ralph Droms, Nancy Griffeth, and Nancy Lynch. The DHCP Failover Protocol: A Formal Perspective. 27th IFIP WG 6.1 International Conference on Formal Methods for Networked and Distributed Systems (FORTE 2007), Tallinn, Estonia, June 26-29, 2007.

We present a formal specification and analysis of a fault-tolerant DHCP algorithm, used to automatically configure certain host parameters in an IP network. Our algorithm uses ideas from an algorithm presented in [DKS+03], but is considerably simpler and at the same time more structured and rigorous. We specify the assumptions and behavior of our algorithm as traces of Timed Input/Output Automata, and prove its correctness using this formalism. Our algorithm is based on a composition of independent subalgorithms solving variants of the classical leader election and shared register problems in distributed computing. The modularity of our algorithm facilitates its understanding and analysis, and can also aid in optimizing the algorithm or proving lower bounds. Our work demonstrates that formal methods can be feasibly applied to complex real-world problems to improve and simplify their solutions.

Shinya Umeno and Nancy Lynch. Proving safety properties of an aircraft landing protocol using I/O Automata and the PVS theorem prover: a case study. FM 2006: Formal Methods, International Symposium of Formal Methods Europe, Hamilton, Ontario Canada, August, 2006. Volume 4085 of Lecture Notes in Computer Science, pages 64-80, Springer, 2006.

Abstract. This paper presents an assertional-style verification of the aircraft landing protocol of

NASA's SATS (Small Aircraft Transportation System) concept [AJCWA] using the I/O automata framework and the PVS theorem prover. We reconstructed the mathematical model of the landing protocol presented in [DMC] as an I/O automaton. In addition, we translated the I/O automaton into a corresponding PVS specification, and conducted a verification of the safety properties of the protocol using the assertional proof technique and the PVS theorem prover.

Shinya Umeno. Proving safety properties of an aircraft landing protocol using timed and untimed I/O automata: a case study. Masters Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, October 2006.

Abstract: This thesis presents an assertional-style verification of the aircraft landing protocol of NASA's SATS (Small Aircraft Transportation System) concept of operation [AJCWA] using the timed and untimed I/O automata frameworks. We construct two mathematical models of the landing protocol using the above stated frameworks. First, we study a discrete model of the protocol, in which the airspace of the airport and every movement of the aircraft are all discretized. The model is constructed by reconstructing a mathematical model presented in [DMC] using the untimed I/O automata framework. Using this model, we verify the safe separation of aircraft in terms of the bounds on the numbers of aircraft in specific discretized areas. In addition, we translate this I/O automaton model into a corresponding PVS specification, and conduct a machine verification of the proof using the PVS theorem prover. Second, we construct a continuous model of the protocol by extending the discrete model using the timed I/O automata framework [KLSV]. A refinement technique has been developed to reason about the external behavior between two systems. We present a new refinement proof technique, a weak refinement using a step invariant. Using this new refinement, we carry over the verification results for the discrete model to the new model, and thus guarantee that the safe separation of aircraft verified for the discrete model also holds for the new model. We also prove properties specific to the new model, such as a lower bound on the spacing of aircraft in a specific area of the airport, using an invariant-proof technique.

Shinya Umeno and Nancy Lynch. Safety Verification of an Aircraft Landing Protocol: A Refinement Approach. Hybrid Systems: Computation and Control (HSCC 2007), Pisa, Italy, April 3-5, 2007, volume 4416 of Lecture Notes in Computer Science, pages 557-572.

Abstract. In this paper, we propose a new approach for formal verification of hybrid systems. To do so, we present a new refinement proof technique, a weak refinement using step invariants. As a case study of the approach, we conduct formal verification of the safety properties of NASA's Small Aircraft Transportation System (SATS) landing protocol. A new model is presented using the timed I/O automata (TIOA) framework [KLSV], and key safety properties are verified. Using the new refinement technique presented in the paper, we first carry over the safety verification results from the previous discrete model studied in [UL] to the new model. We also present properties specific to the new model, such as lower bounds on the spacing of aircraft in specific areas of the airspace.

Ling Cheung, Sayan Mitra and Olivier Pereira. Verifying Statistical Zero Knowledge with Approximate Implementations. Submitted for publication. Available as Cryptology ePrint Archive Report 2007/195.

Abstract. Statistical zero-knowledge (SZK) properties play an important role in designing cryptographic protocols that enforce honest behavior while maintaining privacy. This paper presents a novel approach for verifying SZK properties, using recently developed techniques based on approximate simulation relations. We formulate statistical indistinguishability as an implementation relation in the Task-PIOA framework, which allows us to express computational restrictions. The implementation relation is then proven using approximate simulation relations. This technique separates proof obligations into two categories: those requiring probabilistic reasoning, as well as those that do not. The latter is a good candidate for mechanization. We illustrate the general method by verifying the SZK property of the well-known identification protocol proposed by Girault, Poupard and Stern.

4.7.6 Modeling and Analysis Tools

Members of our group collaborate in the Tempo-TIOA Toolset development effort. The actual development is centered at the University of Connecticut. We do not discuss this work in this final report, since this work has been funded by other AFOSR contracts.

We did carry out one other piece of work on tools, involving development of PVS strategies for proving abstraction properties relating I/O Automata:

Sayan Mitra and Myla Archer. Reusable PVS Proof Strategies for Proving Abstraction Properties of I/O Automata. STRATEGIES 2004, IJCAR Workshop on Strategies in Automated Deduction, Cork Ireland, July 2004.

Abstract: Recent modifications to PVS support a new technique for defining abstraction properties relating automata in a clean and uniform way. This definition technique employs specification templates that can support development of generic high level PVS strategies that set up the standard subgoals of these abstraction proofs and then execute the standard initial proof steps for these subgoals. In this paper, we describe an abstraction specification technique and associated abstraction proof strategies we are developing for I/O automata. The new strategies can be used together with existing strategies in the TAME (Timed Automata Modeling Environment) interface to PVS; thus, our new templates and strategies provide an extension to TAME for proofs of abstraction. We illustrate how the extended set of TAME templates and strategies can be used to prove example I/O automata abstraction properties taken from the literature.

Sayan Mitra and Myla Archer. PVS Strategies for proving abstraction properties automata. In Electronic Notes in Theoretical Computer Science, volume 125(2), 2005, pages 45-65.

Abstract: Abstractions are important in specifying and proving properties of complex systems. To prove that a given automaton implements an abstract specification automaton, one must first find the correct abstraction relation between the states of the automata, and then show that this relation is preserved by all corresponding action sequences of the two automata. This paper describes tool support based on the PVS theorem prover that can help users accomplish the second task, in other words, in proving a candidate abstraction relation correct. This tool support relies on a clean and uniform technique for defining abstraction properties relating automata that uses library theories for defining abstraction relations and templates for specifying automata and abstraction theorems. The paper then describes how the templates and theories allow development of generic, high level PVS strategies that aid in the mechanization of abstraction proofs. These strategies first set up the standard subgoals for the abstraction proofs and then execute the standard initial proof steps for these subgoals, thus making the process of proving abstraction properties in PVS more automated. With suitable supplementary strategies to implement the natural proof steps needed to complete the proofs of any of the standard subgoals remaining to be proved, the abstraction proof strategies can form part of a set of mechanized proof steps that can be used interactively to translate high level proof sketches into PVS proofs. Using timed I/O automata examples taken from the literature, this paper illustrates use of the templates, theories, and strategies described to specify and prove two types of abstraction property: refinement and forward simulation.

4.8 Security Protocols

We began a study of security protocols as an application for our work on probabilistic automata. This has led to a rather in-depth study of ways of modeling certain aspects of security protocols, such as nondeterminism, composability, and real time (in particular, long-lived protocols). We do not discuss this work in detail in this final report, since it has been funded by an NSF ITR contract with Micali and Rivest. Here, we simply list two recent projects that involve new contributions to security protocol design, not just analysis.

Ling Cheung, Joseph A. Cooley, Roger Khazan, and Calvin Newport. Collusion-Resistant Group Key Management Using Attribute-Based Encryption. Proceedings of 1st International Workshop on Group-Oriented Cryptographic Protocols, Wroclaw, Poland, July 2007.

Abstract. This paper illustrates the use of ciphertext-policy attribute-based encryption (CP-ABE), a recently proposed primitive, in the setting of group key management. Specifically, we use the CP-ABE scheme of Bethencourt, Sahai and Waters to implement flat table group key management. Unlike past implementations of flat table, our proposal is resistant to collusion attacks. We also provide efficient mechanisms to refresh user secret keys (for perfect forward secrecy) and to delegate managerial duties to subgroup controllers (for scalability). Finally, we discuss performance issues and directions for future research.

Ling Cheung and Calvin Newport. Provably Secure Ciphertext Policy ABE. Proceedings of the 14th ACM Conference on Computer and Communications Security (CCS), Alexandria, VA, October, 2007. Also, ePrint Report 2007/183, Cryptology ePrint archive, 2007.

Abstract: In ciphertext policy attribute-based encryption (CP-ABE), every secret key is associated with a set of attributes, and every ciphertext is associated with an access structure on attributes. Decryption is enabled if and only if the user's attribute set satisfies the ciphertext access structure. This provides fine-grained access control on shared data in many practical settings, including secure databases and secure multicast.

In this paper, we study CP-ABE schemes in which access structures are AND gates on positive and negative attributes. Our basic scheme is proven to be chosen plaintext (CPA) secure under the decisional bilinear Diffie-Hellman (DBDH) assumption. We then apply the Canetti-Halevi-Katz technique to obtain a chosen ciphertext (CCA) secure extension using one-time signatures. The security proof is a reduction to the DBDH assumption and the strong existential unforgeability of the signature primitive.

In addition, we introduce hierarchical attributes to optimize our basic scheme, reducing both ciphertext size and encryption/decryption time while maintaining CPA security. Finally, we propose an extension in which access policies are arbitrary threshold trees, and we conclude with a discussion of practical applications of CP-ABE.